ABSTRACT

This paper describes an attempt to extract multiple peripheral features of a point \( x(t, f) \) on a time-spectrum (TS) pattern by observing \( n \times n \) neighborhoods of the point, and to incorporate these peripheral features (MPFPs: multiple peripheral feature planes) into the feature extractor of a speech recognition system together with MFCC parameters. Two types of peripheral feature extractor, MPFP-KL and MPFP-LR, are proposed. MPFP-KL adopts the orthogonal bases extracted directly from speech data by using KLT of 7×7 blocks on TS patterns. In MPFP-LR, the upper two primal bases are selected and simplified in the form of \( \Delta \)-operator and \( \Delta \)-operator obtained by linear regression calculation. MPFP-KL and MPFP-LR show significant improvements in comparison with the standard MFCC feature extractor in experiments with the HMM-based ASR system.

1. INTRODUCTION

Time-spectrum (TS) pattern \( x(t, f) \) has long been used for acoustic features in automatic speech recognition (ASR), and recently, dynamic features such as \( \Delta \)-cepstrum, \( \Delta \)-power, etc. have been introduced into ASR [1],[2] and the set of MFCC and dynamic features is widely used. Dynamic features represent peripheral features of a point on a TS pattern \( x(t, f) \) along the time axis, however, we can obtain more information from \( n \times n \) neighborhoods of the point. In this paper, we investigate primal peripheral features embedded in \( n \times n \) blocks of TS patterns first.

In the previous work [3], the feature extraction method based on multiple acoustic feature planes (MAFPs) was applied to a phonetic segment classification task and showed that the method significantly improved the error rate. In the method, a set \( X \) with elements \( x(t, f) \) is mapped onto multiple AFPs (acoustic-feature planes) \( Y_{m} = y_{m}(t, f), m=1, 2, \ldots M \) by using local mapping operators \( \{ G_{m} \} \).

\[
(G_{m} \in G): \quad G_{m} : \quad X \rightarrow Y_{m}
\]

(1)

Firstly, to obtain the \( G_{m} \) the 3×3 orthogonal basis \( \{ \Phi_{1}, \Phi_{2}, \ldots, \Phi_{9} \} \) on TS patterns was extracted directly from a speech database. Then, the orthogonal basis was replaced to a set of modeled operators that simplified \( \{ \Phi_{m} \} \) and made them symmetrical. In this paper, we try not to extract multiple local feature planes (MLFPs) but to extract multiple peripheral feature planes (MPFPs), then incorporate MPFPs into an HMM-based ASR system together with MFCC parameters through the same approach used in the MLFP extraction process [3].

This paper is organized as follows: Section 2 discusses the geometrical structure of 7×7 blocks on TS patterns. Section 3 then outlines methods of implementing the peripheral features in a feature extractor of an ASR system together with MFCC parameters. Finally, Section 4 gives the experimental setup, the results and discussion.

2. OBSERVING PERIPHERAL FEATURES ON TS PATTERNS

We can observe many types of geometrical structures on TS patterns. Figure 1 shows the upper nine elements of an orthogonal basis of 7×7 blocks on TS patterns analyzed by a BPF bank with 24 channels. The 7×7 orthogonal basis was extracted by using Karhunen-Loeve transform (KLT) from speech data described in section 4.1.

From a space-operational point of view, \( \Phi_{1} \) is considered to be a smoothing operator and this neutral operator generally has no effect on feature extraction for ASR. \( \Phi_{2} \) and \( \Phi_{3} \) are the first-order derivative operators with respect to the time axis (\( \Delta_{t} \)-operator) and frequency axis (\( \Delta_{f} \)-operator), respectively. \( \Phi_{4}, \Phi_{9} \) are the second-order derivative operators with respect to the time axis (\( \Delta_{t} \Delta_{t} \)-operator) and frequency axis (\( \Delta_{f} \Delta_{f} \)-operator), respectively, and \( \Phi_{5}, \Phi_{6}, \Phi_{7}, \Phi_{8} \) are...
subspaces that represent ridges and/or valleys on TS patterns.

Time-frequency space operators \(\{F_m\}\), or mapping operators \(\{G_m\}\), map a TS pattern \(x(t, f)\) onto multiple PFPs (peripheral feature planes) \(Y_m = y_m(t, f), m=1, 2, \ldots, M\). An element \(y_m(t, f)\) of MPFP is calculated with 7x7 neighborhoods of \(x(t, f)\) and \(G_m = g_m(t, f)\) by the following equation:

\[
y_m(t, f) = \sum_{i=-3}^{3} \sum_{j=-3}^{3} x(t+i, f+j) \cdot g_m(i, j)
\]

Figure 2 shows an example of the upper three PFPs of an utterance [kaden’tsa] (cadence). In the figure, (A) is an original TS pattern and (B), (C), and (D) represent the 2nd-PFP mapped with a \(\Delta_t\)-operator \(F_2(G_2)\), the 3rd-PFP mapped with a \(\Delta_r\)-operator \(F_3(G_3)\), and the 4th-PFP mapped with \(\Delta_t\Delta_r\)-operator \(F_4(G_4)\), respectively. A positive sign of \(y_m(t, f)\) means a positive slope, a negative sign a negative slope. For example, a clear spectral peak in steady sound is represented by a pair of positive and negative values in the 3rd-PFP. In the figure, patterns on PFP are displayed with absolute values.

3. COMBINING PERIFERAL FEATURES WITH MFCC-PARAMETERS

This chapter describes the methods of extracting peripheral features and combining them with MFCC-parameters in a feature extractor. Figure 3-A shows a standard feature parameters used in current HMM-

![Figure 3-A MFCC with dynamic features](baseline)
based ASR systems. In the feature extractor, an input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segments is applied every 10 ms. The resultant FFT power spectrum is then integrated into the output of 24ch-BPFs that have mel-scaled center-frequencies. Then, 38 feature parameters including 12 static parameters (mel-cepstrum), ΔP (logarithmic power), ΔΔP, and 24 dynamic features (Δt, ΔtΔt) are extracted after converting the output of BPFs into cepstrum coefficients (MFCCs).

Figure 3-B shows the procedure of extracting peripheral features. In the figure, firstly, an output of 24ch-BPF bank x(t, f) is mapped onto PFPs y_m(t, f), m=1,2,…,8; f=1,2,…20 by equation (2). Next, each PFP is converted into cepstrum coefficients c(m,q), m=1,2,..,8; q=1,2,…10 by DCT. Finally, c(m,q) with 80 dimensions are compressed into a selected peripheral-feature-vector z(k), k=1,2,…,24 through KLT by the following equation:

$$z(k) = \sum_{m=1}^{8} \sum_{q=1}^{10} c(m,q) \phi_k(m,q) \quad k=1,2,…,24$$  \hspace{1cm} (3)

where, $\phi_k(m,q)$ is the k-th eigen vector set of KLT. 24 peripheral features z(k) are combined with MFCC static features (12 MFCC + ΔP, ΔΔP).

In chapter 2, we investigated the 7×7 orthogonal basis on TS patterns and found that the upper two primal bases were Δt-operator and Δf-operator. On the other hand, the standard feature vector set of MFCC-based parameters did not include Δf-related ones. Figure 3-C shows the other type of peripheral feature representation. In this figure, two space operators that give two peripheral features of Δt- and Δf- cepstrum are simplified in the form of 7×1-block operator (Δt) and 1×7-block operator (Δf), and the derivative operation is replaced by the calculation of linear regression. 24 peripheral features including 12 Δt-cepstrum coefficients, or dynamic features, are combined with MFCC static features (12 MFCC + ΔP, ΔΔP).

4. EXPERIMENTS

4.1 Speech Database

The following four data sets were used.

D1. Acoustic model design set: A subset of “ASJ (Acoustic Society of Japan) Continuous Speech Database”, consisting of 4,503 sentences uttered by 30 male speakers (16kHz, 16 bit).

D2. Test data set: A subset of “Tohoku University and Matsushita Spoken Word Database”, consisting of 100 words uttered by 10 unknown male speakers.
sampling rate was converted from 24 kHz to 16 kHz.

D3. 7x7 orthogonal basis design data set: A subset of
“ASJ News Corpus (ASJ-JNUS)”, consisting of 2,662
sentences uttered by 53 male speakers.

D4. Eigen-vectors design set for KLT: A subset of the
same ASJ-JNUS corpus used for designing the
orthogonal basis, consisting of 5,569 different
sentences uttered by the same 53 male speakers.

4.2 Experimental Setup

Table 1 shows specifications of the three methods
applied in the feature extractor of the experimental
ASR system. All the methods use MFCC (12), ΔP (1), and ΔΔP (1) and have the same 38 dimensions.
Firstly, 43 Japanese monophone-HMMs with five
states and three loops are designed with a D1 data set.
In the HMM, output probabilities are represented in the form of a Gaussian mixture and covariance matrices
are diagonalized. Next, speaker-independent word-
recognition tests are carried out with a D2 data set.

4.3 Results and Discussion

Table 2 compares the word recognition rates
between the three feature extractors. The results show
that the recognition scores of MPFP-KL with
peripheral features are higher than those of the baseline
extractor that has only dynamic features. Especially,
the score for the one-mixture model of MPFP-KL
is significantly higher than that of the baseline. This fact
suggests that, from a robust feature-extracting point
of view, MFCC with peripheral features is superior to
MFCC with dynamic features.

Why does MPFP-KL show high performance? In
the three orthogonal bases shown in Figure 1, Φ₂ and
Φ₄ were already adopted in the baseline extractor as
dynamic features, however, Φ₃ that shows the upper
contribution in KLT was not introduced yet. MPFP-
LR in Table 2 discards ΔtΔt-cepstrum from the
baseline extractor and adds Δf-cepstrum to it
substitutingly. The result in Table 2 shows the
importance of adding dynamics along the frequency axis to the standard MFCC parameter set. Figure 4
shows the experimental results when various types of
features are added to MFCC.

5. CONCLUSION

A framework for incorporating multiple geometric
structures into the feature extractor of ASR systems
was proposed. The design methodology of mapping
operators for extracting peripheral features was given
by observing the orthogonal basis of speech and by
incorporating primal components into a feature
extractor in a simplified form. The proposed method
based on MPFP-KL or MPFP-LR showed significant improvements in comparison with the standard MFCC
feature extractor in the experiments with the HMM-
based ASR system. It is very important to add
dynamics along the frequency axis to the standard
MFCC parameter set.

REFERENCES

[1] K. Elenius and M. Blomberg, ”Effect of emphasizing
transitional or stationary parts of the speech signal in a

Table 1 Three methods for feature extraction

<table>
<thead>
<tr>
<th>method</th>
<th>parameters</th>
<th>dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>MFCC + Δt + ΔtΔt</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>+ ΔP + ΔΔP</td>
<td></td>
</tr>
<tr>
<td>MPFP-KL</td>
<td>MFCC + KL</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>+ ΔP + ΔΔP</td>
<td></td>
</tr>
<tr>
<td>MPFP- LR</td>
<td>MFCC + Δt + Δt</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>+ ΔP + ΔΔP</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Comparison for three feature extractors

<table>
<thead>
<tr>
<th>method</th>
<th>word correct rate [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mix. = 1</td>
</tr>
<tr>
<td>Baseline</td>
<td>93.58</td>
</tr>
<tr>
<td>MPFP-KL</td>
<td>97.25</td>
</tr>
<tr>
<td>MPFP- LR</td>
<td>97.96</td>
</tr>
</tbody>
</table>

Figure 4 Comparison between MFCC parameter sets.